ChinaXiv合作期刊

RESEARCH PAPER

S-ProvFlow. Storing and Exploring Lineage Data as a Service

Alessandro Spinuso[1†], Malcolm Atkinson[2] & Federica Magnoni[3]

[1]Koninklijk Nederlands Meteorologisch Instituut, De Bilt, Utrecht 3731 GA, The Netherlands

[2]University of Edinburgh, Edinburgh, Edinburgh EH8 9AB, United Kingdom

[3]Istituto Nazionale Geofisica e Vulcanologia, Rome, Lazio 00143, Italy

chinaXiv:202211.00443v1

Keywords: Provenance; Productivity; Workflows; Human-in-the-loop; Visualisation

Citation: Spinuso, A, Atkinson, M., Magnoni, F.: S-ProvFlow. Storing and exploring lineage data as a service. Data Intelligence 4(2), 226-242 (2022). doi: 10.1162/dint_a_00128

Received: July 28, 2021; Revised: December 3, 2021; Accepted: February 4, 2022

ABSTRACT

We present a set of configurable Web service and interactive tools, *s-ProvFlow*, for managing and exploiting records tracking data lineage during workflow runs. It facilitates detailed analysis of single executions. It helps users manage complex tasks by exposing the relationships between data, people, equipment and workflow runs intended to combine productively. Its logical model extends the PROV standard to precisely record parallel data-streaming applications. Its metadata handling encourages users to capture the application context by specifying how application attributes, often using standard vocabularies, should be added. These metadata records immediately help productivity as the interactive tools support their use in selection and bulk operations. Users rapidly appreciate the power of the encoded semantics as they reap the benefits. This improves the quality of provenance for users and management. Which in turn facilitates analysis of collections of runs, enabling users to manage results and validate procedures. It fosters reuse of data and methods and facilitates diagnostic investigations and optimisations. We present S-ProvFlow's use by scientists, research engineers and managers as part of the DARE hyper-platform as they create, validate and use their data-driven scientific workflows.

† Corresponding author: Alessandro Spinuso (Email: alessandro.spinuso@knmi.nl; ORCID: 0000-0002-0077-8491).

ChinaXiv合作期刊

chinaXiv:202211.00443v1

1. INTRODUCTION

The provenance of workflow executions, with its wealth of metadata, needs to be stored, processed, made comprehensible and productively used. The *S-ProvFlow*[①] system supports acquisition and exploration of lineage and provenance data from workflows. It includes a database, a Web service and two interactive tools. Myers *et al.* [1] demonstrated that productive use motivates adoption and improves metadata quality. This improves research objects as their metadata has been refined through use enabling accurate replication—essential for CWFR[②]. Users must be actively engaged through tools that improve their productivity. Much contemporary work enables researchers to author methods abstractly. Those methods are optimally (re) mapped retaining users' specified semantics as technology improves. We focus on the *reverse* information flow, from those evolving systems back to research developers and application experts. We satisfy similar demands for simplicity, comprehensibility and stability. This delivers FAIR benefits to all users by facilitating the access and use of the standardised and persistent provenance traces, and the data, workflows and components accessible via those traces. Our provided interactive tools show this potential. This can be exploited by many other tools, systems and workflow technologies. The metadata used is a mix of generic widely agreed terms, such as system and software identities and geo-spatial references, with discipline and community specific terms representing their knowledge infrastructure in which their work is framed by general (often global) hard-won agreements ([2], page xv). As we illustrate, this is essential for grounding the information in terms usable by them, their peers and successors via their working practices and digital systems. There is a corresponding spectrum of persistent identification. The traces once judged as valuable by domain experts can be allocated standard PIDs. The objects and components they touch will be identified in standard ways or in ways established and sustained by a community. The refinement of the metadata leading to fully described products and experiments can be conducted incrementally, producing updated and refined digital objects. To achieve this, the interactions occurring between data and workflows at different states of maturity should be evaluated within a collaborative and evolving ecosystem, as we will show in Section 4.1. This prototypes a path that other CWFR elements will need to follow to gain wide adoption incrementally in the research communities that have substantial intellectual, technical and political investments already.

Provenance standards are a lingua franca encoding information gathered from multiple layers of a computational workflow. Our comprehensive architecture encompasses generation, management and access to provenance data. In previous work [3, 4], we have already addressed how research developers tune the generation of provenance by injecting metadata instruction in the workflow's operators. This is achieved via an *Active* provenance framework that allows customisation of fine-grained lineage, fostering interaction between users and a workflow's provenance mechanisms [1]. The framework has been demonstrated in the context of a general purpose analysis library for data-streaming pipelines, `dispel4py` [5, 6]. Depending on the requirements, users can instruct a `dispel4py` workflow to extract metadata according to a kernel of

① https://gitlab.com/project-dare/s-ProvFlow
② https://fairdo.org/wg/fdo-cwfr/

ChinaXiv合作期刊

chinaXiv:202211.00443v1

agreed terms specified by initiatives within their domain, as we mention in Section 4, as well as experimental ones. These will be injected into the lineage, alongside those general purpose attributes that characterise the provenance model. In this follow-up paper we present instead a lineage web service and tools that develop understanding and fluency by delivering *immediate* benefits. We show this being used for active monitoring and retrospectively for diverse purposes.

S-ProvFlow exploits standard models [4, 3] to efficiently manage metadata thereby fostering comprehension, usability and *interoperability*. Parameters, intermediate outputs and final results are automatically annotated making the traces and results discoverable and actionable, thereby offering drill-down and FAIR access to a workflow's outcome and enactment history. The system is the provenance service of the DARE[3] platform [7, 8]. It was first released during the VERCE project[4]. In this paper we introduce its web API and the interactive tools. Thanks to the management of the customised provenance traces produced by the excution of the workflow, the interactive tools deliver understanding and control of "live" processes and encourage sharingand reuse of data and methods. We have demonstrated the practical value of provenance when evaluating the basis for evidence and for managing large numbers of runs. *S-ProvFlow* has given us eight years of experience working with domain experts, research developers and the teams who support their use of data and sophisticated computation. We draw on this experience to identify vital requirements for CWFR. These are essential for quality and sustainability of these multi-disciplinary professional collaborations that yield decision support we all depend on.

2. RELATED WORK

Methods for querying and visualising provenance have been developed [9, 10, 11], addressing specific scenarios, workflow systems, or more general mechanisms, such as *ProvStore* [12].Their storage technologies are similarly specialised, e.g., to represent Directed Acyclic Graphs (DAG), *PBase* used Neo4j[5] for the ProvONE model. It enables queries on workflows' traces thatwere previously uploaded onto the system in VisTrails XML format. The interrogations include lineage and execution queries focusing on the involvement of processes within runs and on the relationships between data and processes. For our work we chose the well established document-store, *MongoDB* [13], to give priority to use cases that access the provenance information using data properties and process parameters. This enabled *S-ProvFlow*'s discovery functionalities to exploit the lineage produced via the *Active* framework [3], reflecting each user's context and their metadata. We took on challenges in [14, 15] also addressed by [10], where provenance is annotated with rich metadata and configuration parameters that rely on flexible vocabularies.

The interactive access to provenance data is determined by the quality of visualisation. *MapOrbiter* [16], for instance, summarises the DAG which can be expanded on demand. *S-ProvFlow*, instead, starts with a partial

[3] http://www.project-dare.eu
[4] http://www.verce.eu
[5] http://neo4j.org

ChinaXiv合作期刊

visualisation of the lineage graph. It can be searched or browsed to show process outputs, allowing users to expand and navigate the derivation graph interactively. Alternative techniques, such as the *Sankey* diagram used by *PROV-O-Viz* [17] represent the magnitude of flows between activities. Others map the provenance graph to radial diagrams [18]. This technique, used for the analysis of parallel I/O [19], is also used by *InProv* [11], pioneering its adoption for provenance visualisation for recordings obtained by PASS (Provenance-aware Storage System) [20, 21]. Recently this has been combined with computer graphics to improve the visual efficiency [22]. It reduces visual clutter by bringing the most important nodes to the front. *ProvStore* [12, 23] offers radial diagrams to represent provenance relationships. In *S-ProvFlow* users compose queries to focus on specific metadata terms and values, customising views for single computations or for multiple users and runs. We believe that platforms that aim at the publication and reproducibility of research artifacts, such as *Whole Tale* [24], could benefit from the adoption of *S-ProvFlow*. This would complement their capability to manage preconfigured computing environments, where users re-execute the workflows, with services tomonitor, review and better represent the results. For instance, by performing in *S-ProvFlow* keyprovenance queries that select and render portions of the outputs of the overall workflow.

3. S-PROVFLOW: ARCHITECTURE AND COMPONENTS

S-ProvFlow has multiple components delivering a comprehensive provenance infrastructure. It provides a Web API and tools for interactive exploration of lineage data by users (Figure 1). The underlying provenance model S-PROV[6] [3] builds on the PROV[7] and ProvONE[8] recommendations. This interoperable representation adds elements to encode complex lineage patterns to include process delegation, distribution and statefulness of the workflow operators. In the recent deployment for the DARE platform[9] [25], the components are organised as micro-services optimised by decoupling via message queues, delivering resilience to failures and support for authentication infrastructures.

[6] http://purl.org/s-prov-v1-dev
[7] http://www.w3.org/TR/prov-dm/
[8] https://purl.dataone.org/provone-v1-dev
[9] https://gitlab.com/project-dare/s-ProvFlow/-/tree/master/docker

chinaXiv:202211.00443v1

ChinaXiv合作期刊

chinaXiv:202211.00443v1

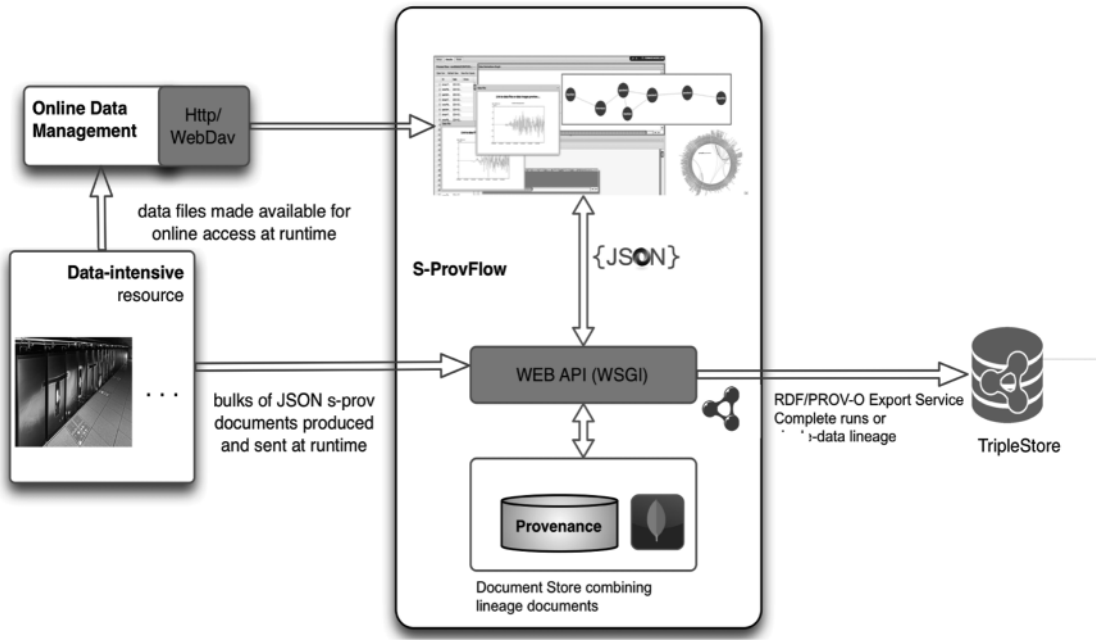

Figure 1. Schematic architecture exploiting the *S-ProvFlow* system for lineage acquisition and exploration. The graphical interface supports direct access to result data, while the API offers an export service to disseminate provenance information, adopting interoperable formats, to general purpose databases, such as triple stores.

3.1 Lineage API

To facilitate the exploitation of the lineage data, *S-ProvFlow* exposes a set of high-level interrogation methods. We present the use cases and relevant implementation details.

Layered Workflow Activity: The execution of a workflow can be examined at different levels of detail, from high-level views based on the classification and functional grouping of workflow elements down to a single element with multiple processing instances. Clients can easily switch between views. We use the recommended *MongoDB* data denormalisation to exploit its powerful aggregation framework[®]. Every time a process is invoked, it also generates a lineage document. Each document contains the same detailed metadata, e.g., the location of the execution, the characteristics of the software and the role of the process as a component of a higher-level function, which may be composed of more operators. This goes alongside dynamic metadata, such as execution time, data volumes, reference values and domain-metadata as specified for the run or process. When queried, the information is aggregated without joins, obtaining complete processing and functional information. The dynamic data is processed and aggregated to deliver the level of abstraction clients have selected. The more documents that are aggregated, in respect to a property of a process, the lower the granularity of the information presented [4].

® https://docs.mongodb.com/v4.0/aggregation/

ChinaXiv合作期刊

chinaXiv:202211.00443v1

Search: The API enables the discovery of data and experiments by performing searches on metadata terms, as well as on the semantic and functional abstractions characterising processes. Intuitive metadata expressions are used to combine value-lists and value-ranges to search for data and workflow executions. To accelerate metadata searches we uses compound indexes[8] on dictionaries of the form {key:<term>, value:<val>}. This allows us to efficiently query dynamic vocabularies, without hitting the limit of the maximum number of indexes allowed.

Data Lineage: The API allows users to navigate the data derivation graph interactively by specifying how much depth should be retrieved at each step. Also in this case, the denormalised approach to the storage combined with linked lists, to represent the PROV *wasDerivedFrom* relationship between data entities in different documents, allows us to easily obtain information about the processes, while navigating through the data derivations. Clients combine graph traversals with metadata filters to view data whose ancestors' properties match their requirements.

Aggregations: The API provides high-level summary methods to extract comprehensive information about single runs or collections of runs. One method covers processing dynamics of a single run, showing data transfers between processes at a configurable granularity. Another reveals collaborative dynamics across runs, such as data-reuse between workflows, infrastructures and users (see Section 4.1). One last method extrapolates information on the metadata in the archive, summarising their role and occurrence for users and workflows. This is used in interfaces to produce personalised recommendations on the terms that could be used in the queries, as shown in Figure 2. All of these methods depend on the powerful aggregation and map-reduce capabilities of the *MongoDB* technology.

3.2 Interactive Tools

Research-developers and administrators may use the API for different purposes. For instance, while the former may validate and analyse the lineage of their experimental results, the latter may monitor how infrastructures and data are exploited by users and applications. *S-ProvFlow* provides tools tailored for both use cases.

The Monitoring and Validation Visualiser (MVV), assists the users in the fine-grain interpretation of the provenance records in order to understand dependencies. It allows them to select and configure viewpoints by specifiable searches over domain metadata, offering data previews and navigation of data dependency graphs. Detailed run-time diagnostics differentiate between stateless and stateful processes, the latter highlighting data retention by operators such as accumulators and mergers. The visual components of the tool are depicted in Figure 2, showing the lineage of a particular workflow in seismology. Users search for workflow executions and data elements by formulating queries using a simple syntax that facilitates metadata searches over ranges or lists of values. Terms may refer to standard vocabularies or be introduced experimentally to evaluate specific applications. Advanced filters on the search operate upon request,

[8] https://docs.mongodb.com/v4.0/core/index-compound/

ChinaXiv合作期刊

reducing the results based on the metadata values of the ancestors in each data derivation tree. The search results can be investigated interactively browsing through the metadata describing both products and processes. Products can be volatile, thereby only described by their metadata. However, when the provenance is associated with actual resources, the tool will show this by offering users preview and download pop-up functions. Every search is assisted through hints, that are updated via the incremental analysis of the whole provenance archive via one of the API's aggregation methods (Section 3.1).

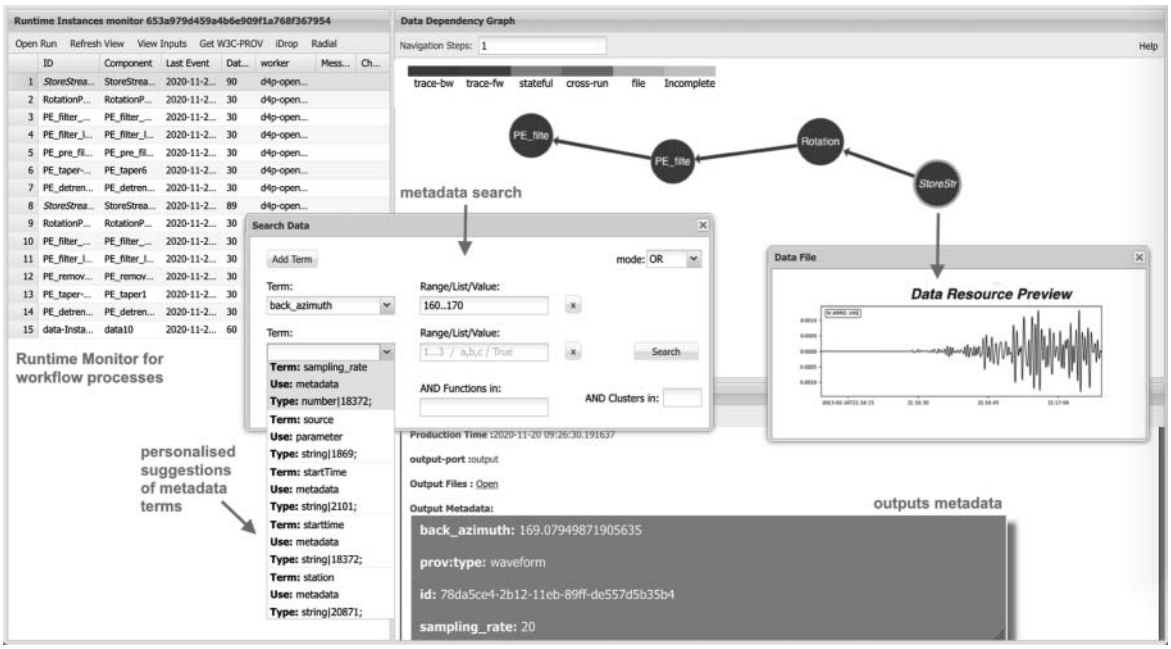

Figure 2. Monitoring and Validation Visualiser combined access to data products and metadata. This example represents the interaction with the provenance information of a run of the *Waveform pre-processing* workflow introduced in Section 4. The *Runtime Monitor* shows the list of active processes. It reports the quantity of data produced, the working node, system or application messages. In the *Data Dependency* graph the yellow circle indicates that the provenance entity links to a concrete data resource. Runs and products can be searched by using metadata expressions.

The Bulk Dependency Visualiser (BDV) offers broader perspectives on computational characteristics as well as collaborative interactions. It combines radial diagrams with configurable grouping and hierarchical edge bundle techniques [22]. It allows its users to dynamically adjust viewing and grouping controls to uncover aspects of the distribution of the processing. This is obtained thanks to the underlying S-PROV model. Its core provenance data model encompassessystem level information in the context of the particular application and workflow's operator semantics, providing an interoperable representation of the workflow's resource-mapping to a target cluster. The BDV uses these capabilities of the model to aggregate data and accommodate views that provide insights on the workflow's enactment at a configurable level of detail, as shown in Figure 3. Besides the exploration of single executions, the BDV can also produce overviews of

chinaXiv:202211.00443v1

ChinaXiv合作期刊

large experiments that involve more researchers who reuse and exchange data via different workflows, with the progressive refinement of the metadata. This visualisation technique could be applied to evolving FDOs to accommodate views that put in the forefront experiments whose metadata has been incrementally refined by peers, to better characterise the methods and the results that contribute to a particular study. This last use case is discussed in Section 4.1 in the context of the particular seismological application.

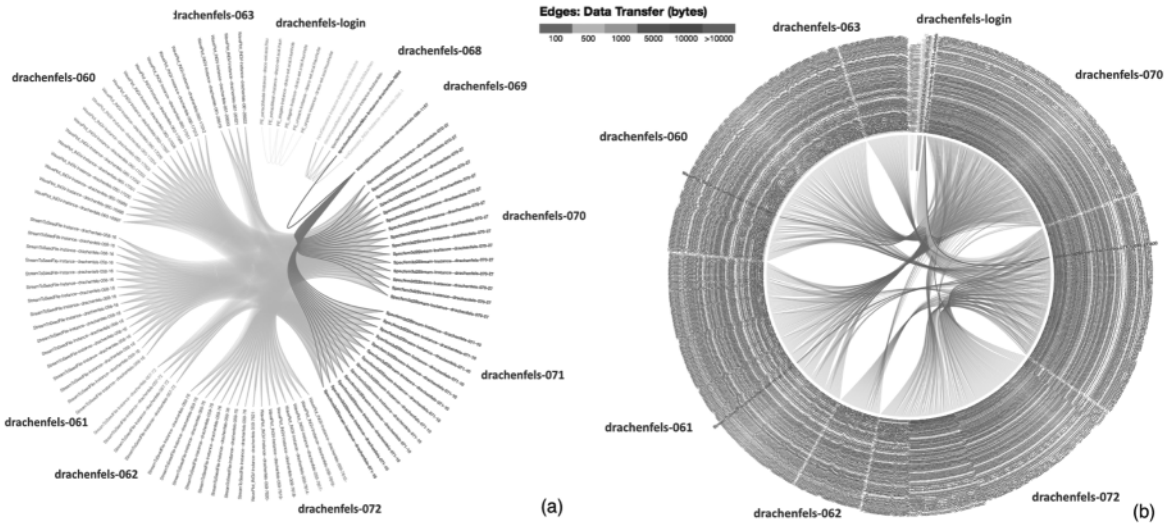

Figure 3. BDV Single workflow visual analytic. Fine-grained radial perspectives for an earthquake simulation workflow. The diagrams indicate overall data transfer between (a) workflows' processes, as well (b) their single invocations in a streaming execution. The colour-coded legend describes the amount of data transferred. The vertices are labeled respectively with processes and invocations ids and are grouped by the computational node they run in the HPC cluster (*Drachenfels* at Fraunhofer SCAI). By hovering on the nodes, incoming (red) and outgoing (green) streams are highlighted.

4. TEST CASE: SEISMIC RAPID ASSESSMENT

Computational seismology is presently facing the challenge of managing increasing amounts of recorded and simulated data, downloaded from rich data archives or simulated by accurate and computationally sophisticated tools. To analyse and exploit these, users need to easily customise the available methods to adroitly explore new opportunities and to meet urgent challenges. Robust provenance-driven tools are needed to smartly organise storage of the data and related metadata, and to encourage their exploration, combination and reuse. This promotes reproducibility and error detection in scientific experiments, and relates well to the general efforts of international organisations handling the wealth of Earth science data[®]. These needs become critically urgent after large seismic events, since reliable and immediate outcomes are fundamental to guide emergency response. In this context, the rapid assessment of seismic ground motion

[®] e.g., http://www.orfeus-eu.org/data/eida; https://www.fdsn.org/; http://ds.iris.edu/ds/

chinaXiv:202211.00443v1

ChinaXiv合作期刊

(RA) is a key application in computational seismology, exposing requirements that embody all the aforementioned needs. It also represents a good example to highlight the applicability of our framework to different scientific contexts (e.g., [26]).

After a large earthquake it is essential to rapidly simulate the propagation of seismic waveforms in surrounding areas and quantitatively estimate specific ground motion parameters to assess the earthquake's impact. Then, comparing and integrating synthetic information with recorded ground motion data improve the understanding of ground response to the earthquake.

The RA theoretical foundations and applicable procedures are well established, and can be exemplified by the high-level steps shown in Figure 4 and detailed in [3, 27, 28]. Some steps are more specific for this seismological test case, some others can be reused (with needed adaptations) in other fields (e.g., volcanology, climate sciences; [26]). All the steps of this application require to be traceable and to have descriptive and explorable metadata in order to support the research scientists in checking, reusing and sharing their methods and findings.

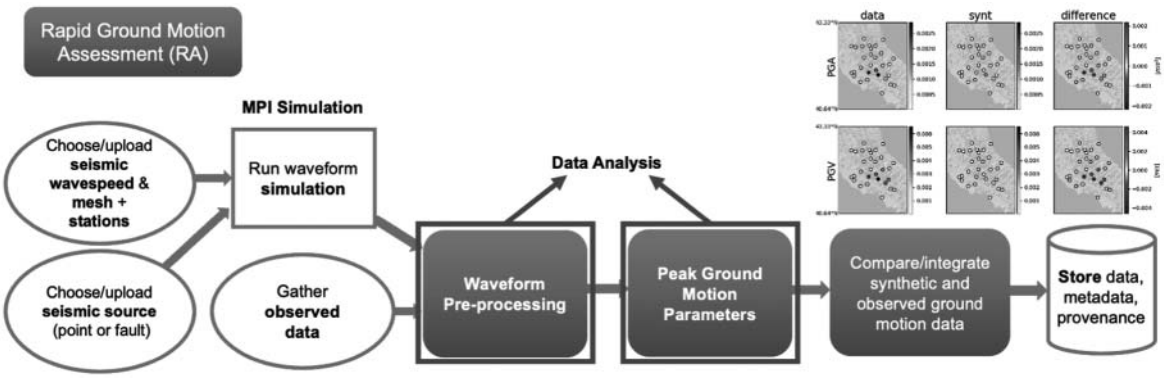

Figure 4. The Rapid Assessment (RA) method analysing the impact of an earthquake composes re-usable tasks that run at different scales and require human supervision and intervention. The image highlights RA data-analysis steps that exploit the benefits of provenance generation and exploitation.

A fundamental step of the RA workflow is the *Waveform pre-processing* stage, which prepares seismological data by making simulated and recorded traces consistent and comparable. It is used in many other seismological applications (with possible variants in the sub-steps) such as inversion for seismic source parameters, seismic tomography and noise cross-correlation analyses (e.g. [29, 30, 31]), and similar processing steps are required for numerous geophysical, and in general scientific, applications (e.g., geodesy, climate sciences, etc.). The system presented in this work guarantees that the results have associated provenance information about all the preprocessing steps (with related parameterisations) that they went through (Figure 2), including data properties after each step. These enable the detection of errors in analyses, and the comparison of different ways of preparing the data and of their effects on the produced results such as the ground motion parameter estimates.

chinaXiv:202211.00443v1

ChinaXiv合作期刊

chinaXiv:202211.00443v1

After pre-processing, the RA analysis requires the extraction of the ground motion parameters from both the synthetic and recorded seismograms for subsequent comparison. In Figure 5 we present the *PGM* (Peak Ground Motion) *Parameters* workflow for a single seismic station and channel. Our implementation applies the combined analysis in parallel on both synthetic and observed data, keeping track of the fine-grain steps they went through and of the acquired metadata. This allows researchers to trace the processing back in case of an error, to combine and compare large amounts of data, and to discover intermediate step results.

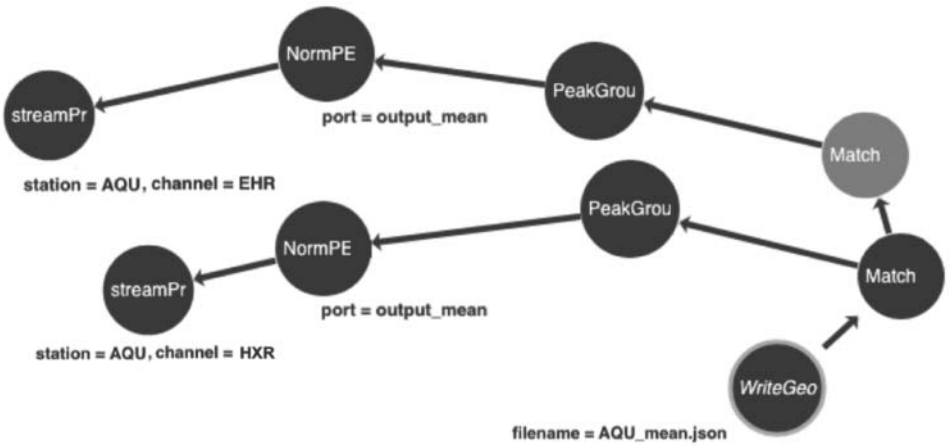

Figure 5. Lineage Precision: The image shows the lineage of a file *AQU mean.json* (yellow circle) produced by the *WriteGeoJSON* process as a result of the PGM workflow. The *wasDerivedFrom* relationships (arrows) show that the file was correctly derived from the mean norm values of the observed and synthetic data channels of the same seismic station AQU, respectively, EHR and HXR. The light blue circle indicates a *stateful derivation*, revealing that the *Match* operator had stored the input data produced by a *PeakGroundMotion* process in its internal state, before matching the data.

Drawing on a strong scientific background, the *Waveform pre-processing* and *PGM Parameters* workflows benefit significantly from *S-ProvFlow*, by having intermediate results properly managed and described by usable metadata. These are typical seismological metadata, which adhere to the recognized standards in the field and are also linkable to well-established infrastructures in the wider Earth science field (e.g., [32, 33], EIDA-ORFEUS[®], FDSN[®], IRIS[®]).

Moreover, there are new, user-customized metadata learned from the workflow executions, providing a database continuously enriched and up-to-date. Users can thus check the results of eachstep even if outputs are not stored. The results and executed steps are traceable in the context of the generating process and

® http://www.orfeus-eu.org/data/eida/
® https://www.fdsn.org/
® http://ds.iris.edu/ds/

ChinaXiv合作期刊

workflow. This fosters their retrieval to allow their comparison and reuse, and to diagnose, validate and fine-tune new experimental analyses. All this makes our metadata and provenance management approach deeply customizable for specific applications and, at the same time, easily adaptable to a multiplicity of scientific workflows in diverse scientific fields, also for operational scenarios [28].

4.1 Collaborative Interactions and Metadata Refinements

The workflow introduced above involves scenarios where collaborative interactions are established between seismologists, by means of reusing data between different stages of the analysis. Figure 6 shows a radial diagram that displays runs executed by two users. The right half of the diagram shows interlinked workflows organised into separated radiants, according to their conceptual tasks. These were described by specifying concepts and metadata to contextualise the methods involved. In contrast, the left side has a poor conceptual characterisation, typical of the early phase of exploration, that results in chaotic and harder to visually analyse provenance graphs. The customisation of the view based on metadata allows users to visually restrict the scope to particular data properties, suggesting, for instance, the reuse of the results of some runs or, depending on the circumstance, the discard of those that show little contribution. The combination of connections and colours make all these different interactions evident. Interactively adjusting viewing parameters with immediate feedback enables users to tune the display until they see what they are looking for. Such visualisation suggests ways to help communities and research managers of computational infrastructures obtain an immediate overview of the interactions across different sites. Especially in the context of CWFR framework, it also shows how the metadata used in a study have been extended and refined over time, towards final and fully described experiments. These comprehensive diagrams are a significant help when coordinating a long-running or extensive research campaign. They may be used for public outreach and within official reports, providing a sensible and tangible perception of the actual exploitation of a distributed data-intensive platform, serving yet another category of consumers through the same underlying provenance model.

chinaXiv:202211.00443v1

ChinaXiv合作期刊

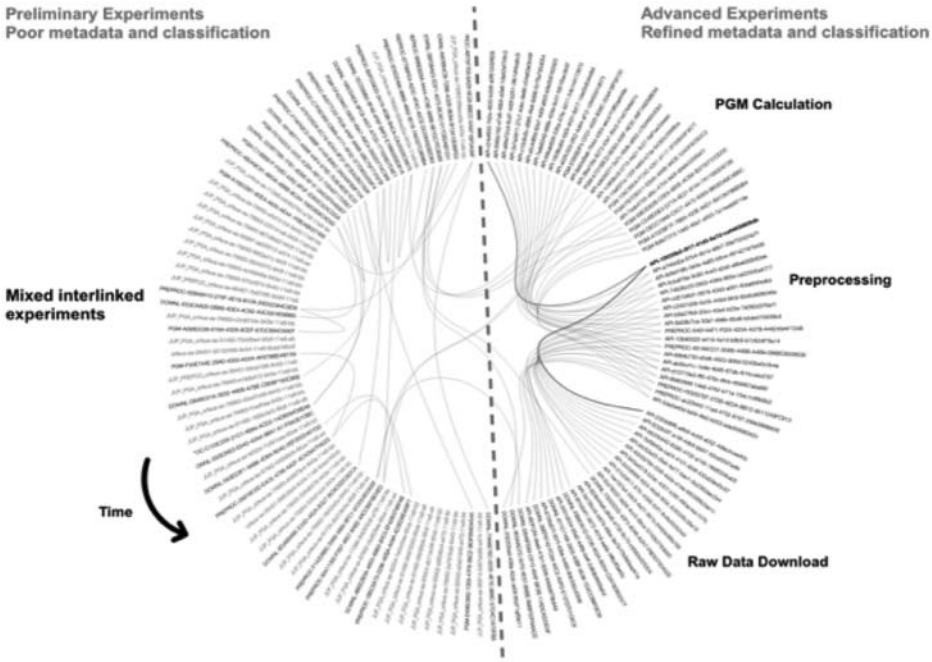

Figure 6. BDV Radial diagram highlighting data reuse between different workflows of the RA use case and refinement of the associated metadata. The runs were selected by interactively querying the provenance archive to find those using a common set of seismic stations. The runs were performed by two different users (blue and black vertex). The right half presents interlinked workflows with a refined contextualisation of the lineage. It shows how the improvements of the metadata applied over time by one of the users, yield more informative content and better visualisation.

5. CONCLUSIONS AND FUTURE WORK

The incremental improvement of the relevance and usefulness of provenance increases confidence in the possibilities for its exploitation, thereby promoting awareness of its importance. The provenance traces themselves are candidate FDOs [34]. They also provide critical actionable information about the workflows, e.g., to assess their validity, usefulness and efficiency, and about each enactment and its data products, all of which may eventually be FDOs, but this depends on stimulating adoption of CWFR standards, which will inevitably be an incremental process following the kinds of path we have pioneered. Through the active participation of the experts, provenance is ready made for validation and results' management use cases. We explored technical solutions and standard models to be applied to the next generation of WMSs [35], which have to encompass the challenges, identified by the FDO®, concerning the evaluation and traceability of experimental results. This depended on co-design and co-development with domain experts in communities that had long-established global knowledge infrastructure with corresponding standards and agreed

® https://fairdo.org/

chinaXiv:202211.00443v1

ChinaXiv合作期刊

practices. In this paper, we have focused on our work with computational seismologists. In a companion paper [36] we report how provenance management empowers thereproducibility of interactive workspaces, beyond workflows, with use cases which also include climate scientists. Edwards [2] makes clear the extensive and complex knowledge infrastructure that has taken over a century to incrementally develop. The standards and technologies we produce have to prove their value alongside those established systems before they will begin to penetrate into their *established* working practices. We consider that improving the quality and use of provenance by delivering immediate benefits to a broad range of users will encourage adoption and engagement. Their improved productivity and the improved quality of decision support will motivate the investment in sustainably collecting and preserving vital provenance records with sufficient content. It will have long-term benefits for the quality of research procedures and the evidence they produce that underpins life-critical decisions. We have adopted services and tools developed around our framework to demonstrate its effectiveness. The developed interactive tools provide easy access, visualization and navigation through the provenance information and metadata, also offering direct links to physical data resources. Researchers exploit these tools to improve their science, by quickly detecting and solving anomalies, and optimizing the combination of multiple runs and data for complex applications. Research engineers and developers are facilitated in improving the resource and data exploitation (Section 4.1.) We illustrated many ofthese capabilities in relation to the use cases of a real application in seismology. In future work, we want to address and improve the preliminary results [25] of enabling the import of *CWLProv* into *s-ProvFlow*, in order to scale the benefit of *S-ProvFlow* to a wider collection of WMSs.

Similar interdisciplinary co-development demonstrating immediate benefits will test and develop the other aspects of the canonical workflow technologies and incentivise widespread experimental adoption leading to sustained growth in quality, capabilities and adoption. We therefore anticipate collaborations embedding in many research contexts to improve the standard including the provenance, developing the tools and work environments it enables, to build momentum for its adoption. This will deliver more FDOs corresponding to all aspects of the supported research and extend the use of FDO standards and representations deeper into the established practices.

AUTHOR CONTRIBUTIONS

A. Spinuso (spinuso@knmi.nl) and M. Atkinson (Malcolm.Atkinson@ed.ac.uk) contributed to the research, design and implementation of the system. F. Magnoni (federica.magnoni@ingv.it) verified its capabilities by implementing the use case that took advantage ofthe framework. All authors contributed to the writing of the manuscript.

ACKNOWLEDGEMENTS

This work was supported by the EU FP7-Infrastructure project VERCE (No. 283543) and EU H2020 project DARE (No. 777413). These projects are composed of large teams of software engineers and researchers in Climate and Earth Sciences, who contributed to the implementation and adoption of the technologies illustrated by this work. We thank them for their continuous support and proactive participation.

chinaXiv:202211.00443v1

ChinaXiv合作期刊

chinaXiv:202211.00443v1

REFERENCES

[1] Myers, J., et al.: Towards sustainable curation and preservation: The sead project's data services approach. In: 2015 IEEE 11th International Conference on e-Science, pp. 485–494 (2015)

[2] Edwards, P.N.: A vast machine: Computer models, climate data, and the politics of global warming. MIT Press, Cambridge (2010)

[3] Spinuso, A., Atkinson, M., Magnoni, F.: Active provenance for data-intensive workflows: Engaging users and developers. In: 2019 15th International Conference on eScience (eScience), pp. 560–569 (2019)

[4] Spinuso, A.: Active provenance for data intensive research. PhD dissertation, The University of Edinburgh (2018). Available at: http://hdl.handle.net/1842/33181. Accessed 4 February 2022

[5] Filgueira, R., et al.: dispel4py: An agile framework for data-intensive escience. In: The 11th IEEE International Conference on e-Science, pp. 454–464 (2015)

[6] Filguiera, R., et al.: dispel4py: A python framework for data-intensive scientific computing. In 2014 International Workshop on Data Intensive Scalable Computing Systems, pp. 9–16 (2014)

[7] Klampanos, I., et al.: Dare: A reflective platform designed to enable agile data-driven research on the cloud. In: The 15th International Conference on eScience (eScience), No. 19473092 (2019)

[8] Atkinson, M., et al: Comprehensible control for researchers and developers facing data challenges. In: The 15th International Conference on eScience (eScience), No. 19473140 (2019)

[9] Hunter, J., Cheung, K.: Provenance explorer—A graphical interface for constructing scientific publication packages from provenance trails. International Journal on Digital Libraries 7, 99–107(2007)

[10] Cuevas-Vicenttín, V., et al.: The PBase scientific workflow provenance repository. International Journal of Digital Curation 9(2), 28–38 (2014)

[11] Borkin, M.A., et al.: Evaluation of filesystem provenance visualization tools. IEEE Transactions on Visualization and Computer Graphics 19(12), 2476–2485 (2013)

[12] Huynh, T.D., Moreau, L.: Provstore: A public provenance repository. In: International Provenance and Annotation Workshop, pp. 275–277 (2014)

[13] MongoDB Document-oriented Data Base (2020). Available at: http://mongodb.org. Accessed 1 February 2022

[14] De Oliveira, D., Silva, V., Mattoso, M.: How much domain data should be in provenance databases? In: Proceedings of the 7th USENIX Conference on Theory and Practice of Provenance (TaPP' 15), pp. 1–9 (2015)

[15] Gadelha, L.M.R., et al.: MTCProv: A practical provenance query framework for many-task scientific computing. Distributed Parallel Databases 30, 351–370 (2012)

[16] Seltzer, M., Macko, P.: Provenance map orbiter: Interactive exploration of large provenance graphs. In: Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP'11), pp. 20–21(2011)

[17] Hoekstra, R, Groth, P: Prov-O-Viz—understanding the role of activities in provenance. In: Ludäscher, B., Plale, B. (eds.) Provenance and Annotation of Data and Processes, pp. 215–220. Springer International Publishing, Cham (2015)

[18] Draper, G.M., Livnat, Y., Riesenfeld, R.F.: A survey of radial methods for information visualization. IEEE Transactions on Visualization and Computer Graphics 15(5), 759–776 (2009)

[19] Sigovan, C., et al.: A visual network analysis method for large-scale parallel I/O systems. In: IEEE 27th International Symposium on Parallel Distributed Processing (IPDPS), pp. 308–319 (2013)

[20] PASS: Provenance-aware storage systems. Available at: http://www.eecs.harvard.edu/syrah/pass/. Accessed 1 February 2022

ChinaXiv合作期刊

[21] Muniswamy-Reddy, K.-K., et al.: Provenance-aware storage systems. In: USENIX Annual Technical Conference, General Track, pp. 43–56 (2006)
[22] Holten, D.: Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. IEEE Transactions on Visualization and Computer Graphics 12(5), 741–748 (2006)
[23] ProvStore, provenance storage and distribution. Available at: https://openprovenance.org/store/. Accessed 1 February 2022
[24] Brinckman, A., et al.: Computing environments for reproducibility: Capturing the “whole tale”. Future Generation Computer Systems 94, 854–867 (2019)
[25] Spinuso, A.: D3.4 data lineage services ii (2020). Available at: https://doi.org/10.5281/zenodo.4738814. Accessed 1 February 2022
[26] Malcolm, A., et al.: Dare architecture and technology D2.2 (Version 1) (2020). Available at: https://doi.org/10. 5281/zenodo.4733801. Accessed 1 February 2022
[27] Magnoni, F., et al.: D6.3 pilot tools and services, execution and evaluation report I (2019). Available at: https://doi.org/10.5281/zenodo.4739217. Accessed 1 February 2022
[28] Magnoni, F., et al.: D6.4 pilot tools and services, execution and evaluation report II (2020). Available at: https://doi. org/10.5281/zenodo.4740027. Accessed 1 February 2022
[29] Scognamiglio, L., et al.: Uncertainty estimations for moment tensor inversions: The issue of the 2012 May 20 Emilia earthquake. Geophysical Journal International 206(2), 792–806 (2016)
[30] Bozdağ, E,. et al.: Global adjoint tomography: First-generation model. Geophysical Journal International 207(3), 1739–1766 (2016)
[31] Zaccarelli, L., et al.: Variations of crustal elastic properties during the 2009 L’Aquila earthquake inferred from cross-correlations of ambient seismic noise. Geophysical Research Letters 38(24), L24304 (2011)
[32] Danecek, P., et al.: The Italian node of the European integrated data archive. Seismological Research Letters 92(3), 1726–1737 (2021)
[33] Michelini, A., et al.: INSTANCE—The Italian seismic dataset for machine learning. Earth System Science Data Discuss Preprint (2021). Available at: https://doi.org/10.5194/essd-2021-164. Accessed 1 February 2022
[34] De Smedt, K., Koureas, D., Wittenburg, P.: FAIR digital objects for science: From data pieces to actionable knowledge units. Publication 8(2), Article No. 21 (2020)
[35] Deelman, E., et al.: The future of scientific workflows. The International Journal of High Performance Computing Applications 32(1), 159–175 (2018)
[36] Spinuso, A., et al.: SWIRRL: Managing provenance-aware and reproducible workspaces. Data Intelligence 4(2), 243–258 (2022)

chinaXiv:202211.00443v1

ChinaXiv合作期刊

chinaXiv:202211.00443v1

AUTHOR BIOGRAPHY

Alessandro Spinuso is a researcher at the RD Observations and Data Technology division of the Royal Netherlands Meteorological Institute (KNMI). He earned her Ph.D. in Computer Science at the University of Edinburgh (UK) in 2017. At KNMI, he covers the roles of Researcher and Product Owner within an Agile RD team developing Provenance-aware Data Analysis services. His main research interest is the management and the exploitation of provenance information in the context of user controlled computational environments providing notebooks and workflow systems for data-intensive analysis. He is involved in several EU initiatives (H2020, Copernicus), focusing on the development of e-Science infrastructures for Earth Science research in Europe (EPOS, ENVRIFair, IS-ENES3, DARE, and C3S). More recently, he is an invited expert to the IPCC TG-Data, a working group dedicated to the FAIR management of the data and methods that will be published in the next IPCC reports.
ORCID: 0000-0002-0077-8491

Malcolm Atkinson, Ph.D., FBCS, FRSE, has spent 53 years improving our capacity through education and research to enable individuals, organisation and society to make best use of their data. He leads the data-intensive research group, focusing on socio-technical architectures to address this challenge. He is Professor of e-Science in the Artificial Intelligence Applications Institute, School of Informatics, University of Edinburgh.
ORCID: 0000-0003-2632-0013

ChinaXiv合作期刊

Federica Magnoni is a researcher in computational seismology at Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome, Italy. She earned her Ph.D. in Geophysics at the Univer sity of Bologna (Italy) in 2012. She has substantial experience in 3D seismic wave propagation forward and inverse problems and in ground shaking assessment, exploiting numerical methods for realistic 3D earth models and point or finite earthquake source models. She participated in numerous projects that exploit national and international HPC resources for HPC- and data-intensive applications (PRACE and ISCRA projects). She worked and is presently involved in several EU projects (FP7, H2020), focusing on the development of e-Science infrastructures for Earth Science (VERCE, EPOS-IP, and DARE), on scientific applications with flagship European codes prepared for pre-Exascale and Exascale computations (ChEESE), as well as national projects on earthquake near real-time simulation and role of social media in emergency communication (PRIN 2012—SHAKEnetworks), or on the exploitation of machine learning techniques for seismological applications (Pianeta Dinamico - SOME).
ORCID: 0000-0002-0833-8044

chinaXiv:202211.00443v1